Uniform Convergence for Learning Binary Classifcation

- Given a concept class $\mathcal{C}$, and a training set sampled from $\mathcal{D}$, $\{(x_i, c(x_i)) \mid i = 1, \ldots, m\}$.
- For any $h \in \mathcal{C}$, let $\Delta(c, h)$ be the set of items on which the two classifiers differ: $\Delta(c, h) = \{x \in U \mid h(x) \neq c(x)\}$
- For the realizable case we need a training set (sample) that with probability $1 - \delta$ intersects every set in

$$\{\Delta(c, h) \mid Pr(\Delta(c, h)) \geq \epsilon\} \quad (\epsilon\text{-net})$$

- For the unrealizable case we need a training set that with probability $1 - \delta$ estimates, within additive error ϵ, every set in

$$\Delta(c, h) = \{x \in U \mid h(x) \neq c(x)\} \quad (\epsilon\text{-sample}).$$

- Under what conditions can a finite sample achieve these requirements?
    - What sample size is needed?

Uniform Convergence Sets

Given a collection R of sets in a universe X, under what conditions a finite sample N from an arbitrary distribution $\mathcal{D}$ over X, satisfies with probability $1 - \delta$,

1. $$\forall r \in R, \ \Pr_{\mathcal{D}}(r) \geq \epsilon \Rightarrow \ r \cap N \neq \emptyset \qquad (\epsilon\text{-net})$$

2. for any $r \in R$,

$$\left| \Pr_{\mathcal{D}}(r) - \frac{|N \cap r|}{|N|} \right| \leq \varepsilon \qquad (\epsilon\text{-sample})$$

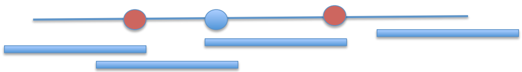
Vapnik–Chervonenkis (VC) - Dimension

(X, R) is called a "range space":

- $X =$ finite or infinite set (the set of objects to learn)
- R is a family of subsets of X, $R \subseteq 2^X$.
- For a finite set $S \subseteq X$, $|S| = m$, define the projection of R on S,
$$\Pi_R(S) = \{r \cap S \mid r \in R\}.$$
- If $|\Pi_R(S)| = 2^m$ we say that R shatters S.
- The VC-dimension of (X, R) is the maximum size of $S \subseteq X$ that is shattered by R. If there is no maximum, the VC-dimension is ∞.

The VC-Dimension of a Collection of Intervals

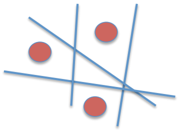
C = collections of intervals in [A,B] – can shatter 2 point but not 3. No interval includes only the two red points

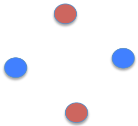


The VC-dimension of *C* is 2

Collection of Half Spaces in the Plane

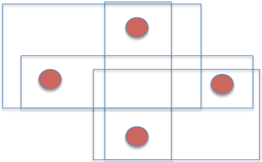
C – all half space partitions in the plane. Any 3 points can be shattered:

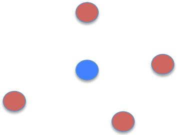




- Cannot partition the red from the blue points
- The VC-dimension of half spaces on the plane is 3
- The VC-dimension of half spaces in d-dimension space is d+1

Axis-parallel rectangles on the plane



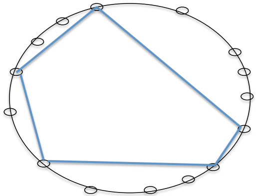


4 points that define a convex hull can be shattered.

No five points can be shattered since one of the points must be in the convex hull of the other four.

Convex Bodies in the Plane

- *C* – all convex bodies on the plane



Any subset of the point can be included in a convex body.
The VC-dimension of *C* is ∞

A Few Examples

- $\mathcal{C} =$ set of intervals on the line. Any two points can be shattered, no three points can be shattered.
- $\mathcal{C} =$ set of linear half spaces in the plane. Any three points can be shattered but no set of 4 points. If the 4 points define a convex hull let one diagonal be 0 and the other diagonal be 1. If one point is in the convex hull of the other three, let the interior point be 1 and the remaining 3 points be 0.
- $\mathcal{C} =$ set of axis-parallel rectangles on the plane. 4 points that define a convex hull can be shattered. No five points can be shattered since one of the points must be in the convex hull of the other four.
- $\mathcal{C} =$ all convex sets in R^2. Let S be a set of n points on a boundary of a cycle. Any subset $Y \subset S$ defines a convex set that doesn't include $S \setminus Y$.

Estimating Probabilities - ε-sample

Definition

An ε-sample for a range space (X, R), with respect to a probability distribution $\mathcal{D}$ defined on X, is a subset $N \subseteq X$ such that, for any $r \in R$,

$$\left| \Pr_{\mathcal{D}}(r) - \frac{|N \cap r|}{|N|} \right| \leq \varepsilon \ .$$

Theorem

Let $(X, \mathcal{R})$ be a range space with VC dimension d and let $\mathcal{D}$ be a probability distribution on X. For any $0 < \epsilon, \delta < 1/2$, there is an

$$m = O\left(\frac{d}{\epsilon^2} \ln \frac{d}{\epsilon} + \frac{1}{\epsilon^2} \ln \frac{1}{\delta} \right)$$

such that a random sample from $\mathcal{D}$ of size greater than or equal to m is an ϵ-sample for X with with probability at least $1 - \delta$.

Sauer's Lemma

For a finite set $S \subseteq X$, $s = |S|$, define the projection of R on S,

$$\Pi_R(S) = \{r \cap S \mid r \in R\}.$$

Theorem

Let (X, R) be a range space with VC-dimension d, for $S \subseteq X$, such that $|S| = m$,

$$|\Pi_R(S)| \leq \sum_{i=0}^{d} \binom{m}{i}.$$

For $m = d$, $|\Pi_R(S)| = 2^d$, and for $m > d \geq 2$, $|\Pi_R(S)| \leq m^d$.

Proof

- By induction on d and (for each d) on n, obvious for $d = 0, 1$ with any n.
- Assume that the claim holds for all $|S'| \leq n - 1$ and $d' \leq d - 1$ and let $|S| = n$.
- Fix $x \in S$ and let $S' = S - \{x\}$.

$$
\begin{aligned}
|\Pi_R(S)| &= |\{r \cap S \mid r \in R\}| \\
|\Pi_R(S')| &= |\{r \cap S' \mid r \in R\}| \\
|\Pi_{R(x)}(S')| &= |\{r \cap S' \mid r \in R \text{ and } x \notin r \text{ and } r \cup \{x\} \in R\}|
\end{aligned}
$$

- For $r_1 \cap S \neq r_2 \cap S$ we have $r_1 \cap S' = r_2 \cap S'$ iff $r_1 = r_2 \cup \{x\}$, or $r_2 = r_1 \cup \{x\}$. Thus,

$$|\Pi_R(S)| = |\Pi_R(S')| + |\Pi_{R(x)}(S')|$$

Fix $x \in S$ and let $S' = S - \{x\}$.

$$\begin{aligned} |\Pi_R(S)| &= |\{r \cap S \mid r \in R\}| \\ |\Pi_R(S')| &= |\{r \cap S' \mid r \in R\}| \\ |\Pi_{R(x)}(S')| &= |\{r \cap S' \mid r \in R \text{ and } x \notin r \text{ and } r \cup \{x\} \in R\}| \end{aligned}$$

- The VC-dimension of $(S, \Pi_R(S))$ is no more than the VC-dimension of (X, R), which is d.
- The VC-dimension of the range space $(S', \Pi_R(S'))$ is no more than the VC-dimension of $(S, \Pi_R(S))$ and $|S'| = n - 1$, thus by the induction hypothesis $|\Pi_R(S')| \leq \sum_{i=0}^{d} \binom{n-1}{i}$.
- For each $r \in \Pi_{R(x)}(S')$ the range set $\Pi_S(R)$ has two sets: r and $r \cup \{x\}$. If B is shattered by $(S', \Pi_{R(x)}(S'))$ then $B \cup \{x\}$ is shattered by (X, R), thus $(S', \Pi_{R(x)}(S'))$ has VC-dimension bounded by $d - 1$, and $|\Pi_{R(x)}(S')| \leq \sum_{i=0}^{d-1} \binom{n-1}{i}$.

$$|\Pi_R(S)| = |\Pi_R(S')| + |\Pi_{R(x)}(S')|$$

$$\begin{aligned} |\Pi_R(S)| &\leq \sum_{i=0}^{d} \binom{n-1}{i} + \sum_{i=0}^{d-1} \binom{n-1}{i} \\ &= 1 + \sum_{i=1}^{d} \left(\binom{n-1}{i} + \binom{n-1}{i-1} \right) \\ &= \sum_{i=0}^{d} \binom{n}{i} \leq \sum_{i=0}^{d} \frac{n^i}{i!} \leq n^d \end{aligned}$$

[We use $\binom{n-1}{i-1} + \binom{n-1}{i} = \frac{(n-1)!}{(i-1)!(n-i-1)!}\left(\frac{1}{n-i} + \frac{1}{i}\right) = \binom{n}{i}$]

The number of distinct concepts on n elements grows polynomially in the VC-dimension!

ε-sample

Definition

An ε-sample for a range space (X, R), with respect to a probability distribution $\mathcal{D}$ defined on X, is a subset $N \subseteq X$ such that, for any $r \in R$,

$$\left| \Pr_{\mathcal{D}}(r) - \frac{|N \cap r|}{|N|} \right| \le \varepsilon \ .$$

Theorem

Let $(X, \mathcal{R})$ be a range space with VC dimension d and let $\mathcal{D}$ be a probability distribution on X. For any $0 < \epsilon, \delta < 1/2$, there is an

$$m = O\left(\frac{d}{\epsilon^2} \ln \frac{d}{\epsilon} + \frac{1}{\epsilon^2} \ln \frac{1}{\delta} \right)$$

such that a random sample from $\mathcal{D}$ of size greater than or equal to m is an ϵ-sample for X with with probability at least $1 - \delta$.

Proof of the ε-Sample Theorem

Let N be a set of m independent samples from X according to $\mathcal{D}$. Let

$$E_1 = \left\{ \exists r \in R \text{ s.t. } \left| \frac{|N \cap r|}{m} - \Pr(r) \right| > \varepsilon \right\}.$$

We want to show that $\Pr(E_1) \leq \delta$.

Choose another set T of m independent samples from X according to $\mathcal{D}$. Let

$$E_2 = \left\{ \exists r \in R \text{ s.t. } \left| \frac{|N \cap r|}{m} - \Pr(r) \right| > \varepsilon \ \wedge \ \left| \Pr(r) - \frac{|T \cap r|}{m} \right| \leq \varepsilon/2 \right\}$$

Lemma

$$\Pr(E_2) \leq \Pr(E_1) \leq 2 \Pr(E_2).$$

Lemma

$\Pr(E_2) \leq \Pr(E_1) \leq 2\Pr(E_2)$.

$$E_1 = \left\{ \exists r \in R \text{ s.t. } \left| \frac{|N \cap r|}{m} - \Pr(r) \right| > \varepsilon \right\}$$

$$E_2 = \left\{ \exists r \in R \text{ s.t. } \left| \frac{|N \cap r|}{m} - \Pr(r) \right| > \varepsilon \ \wedge \ \left| \frac{|T \cap r|}{m} - \Pr(r) \right| \leq \varepsilon/2 \right\}$$

For $m \geq \frac{24}{\varepsilon}$,

$$\begin{aligned} \frac{\Pr(E_2)}{\Pr(E_1)} &= \frac{\Pr(E_1 \cap E_2)}{\Pr(E_1)} = \Pr(E_2|E_1) \geq \Pr(|\frac{|T \cap r|}{m} - \Pr(r)| \leq \varepsilon/2) \\ &\geq 1 - 2e^{-\varepsilon m/12} \geq 1/2 \end{aligned}$$

Instead of bounding the probability of

$$E_2 = \left\{ \exists r \in R \text{ s.t. } \left| \frac{|N \cap r|}{m} - \Pr(r) \right| > \varepsilon \ \wedge \ \left| \frac{|T \cap r|}{m} - \Pr(r) \right| \leq \varepsilon/2 \right\}$$

we bound the probability of

$$E_2' = \{\exists r \in R \mid ||r \cap N| - |r \cap T|| \geq \frac{\epsilon}{2} m\}.$$

Since

$$||r \cap N| - |r \cap T|| \geq ||r \cap N| - m \Pr_{\mathcal{D}}(r)| - ||r \cap T| - m \Pr_{\mathcal{D}}(r)| \geq \frac{\epsilon}{2} m.$$

Lemma

$$\Pr(E_1) \leq 2\Pr(E_2) \leq 2\Pr(E_2') \leq 2(2m)^d e^{-\epsilon^2 m/8}.$$

- Since N and T are random samples, we can first choose a random sample of $2m$ elements $Z = z_1, \ldots, z_{2m}$ and then partition it randomly into two sets of size m each.
- Since Z is a random sample, any partition that is independent of the actual values of the elements generates two random samples.
- We will use the following partition: for each pair of sampled items z_{2i-1} and z_{2i}, $i = 1, \ldots, m$, with probability $1/2$ (independent of other choices) we place z_{2i-1} in T and z_{2i} in N, otherwise we place z_{2i-1} in N and z_{2i} in T.

For $r \in R$, let B_r be the event

$$B_r = \left\{ ||r \cap N| - |r \cap T|| \geq \frac{\varepsilon}{2} m \right\}. \qquad E_2' = \bigcup_{r \in R} B_r$$

The event B_r depends only on the random partition of Z into N and T. Its doesn't depend on the selection of Z.

- If $z_{2i-1}, z_{2i} \in r$ or $z_{2i-1}, z_{2i} \notin r$ they don't contribute to the value of $||r \cap N| - |r \cap T||$.
- If just one of the pair z_{2i-1} and z_{2i} is in r then their contribution is $+1$ or -1 with equal probabilities.
- There are at least $\epsilon m/2$ pairs that contribute $+1$ or -1 with equal probabilities. Applying the Chernoff bound we have

$$Pr(E_r) \leq e^{-\epsilon^2 m/8}.$$

$$Pr(E_r) \leq e^{-\epsilon^2 m/8}.$$

$$E_2' = \{\exists r \in R \mid ||r \cap N| - |r \cap T|| \geq \tfrac{\epsilon}{2} m\} = \bigcup_{r \in R} B_r.$$

Since the projection of R on $T \cup N$ has no more than $(2m)^d$ different ranges, we have

$$\Pr(E_1) \leq 2 \Pr(E_2') \leq 2(2m)^d e^{-\epsilon^2 m/8}.$$

To complete the proof we need to show that for

$$m \geq \frac{32d}{\epsilon^2} \ln \frac{64d}{\epsilon^2} + \frac{16}{\epsilon^2} \ln \frac{1}{\delta}$$

we have

$$(2m)^d e^{-\epsilon^2 m/8} \leq \delta.$$

To complete the proof we show that for

$$m \geq \frac{32d}{\epsilon^2} \ln \frac{64d}{\epsilon^2} + \frac{16}{\epsilon^2} \ln \frac{1}{\delta}$$

we have

$$(2m)^d e^{-\epsilon^2 m/8} \leq \delta.$$

Equivalently, we require

$$\epsilon^2 m/8 \geq \ln(1/\delta) + d \ln(2m).$$

Clearly $\epsilon^2 m/16 \geq \ln(1/\delta)$, since $m > \frac{16}{\epsilon^2} \ln \frac{1}{\delta}$.

To show that $\epsilon^2 m/16 \geq d \ln(2m)$ we use:

Lemma

If $y \geq x \ln x > e$, *then* $\frac{2y}{\ln y} \geq x$.

Proof.

For $y = x \ln x$ we have $\ln y = \ln x + \ln \ln x \leq 2 \ln x$. Thus

$$\frac{2y}{\ln y} \geq \frac{2x \ln x}{2 \ln x} = x.$$

Differentiating $f(y) = \frac{\ln y}{2y}$ we find that $f(y)$ is monotonically decreasing when $y \geq x \ln x \geq e$, and hence $\frac{2y}{\ln y}$ is monotonically increasing on the same interval, proving the lemma. □

Let $y = 2m \geq \frac{64d}{\epsilon^2} \ln \frac{64d}{\epsilon^2}$ and $x = \frac{64d}{\epsilon^2}$, we have $\frac{4m}{\ln(2m)} \geq \frac{64d}{\epsilon^2}$, so $\frac{\epsilon^2 m}{16} \geq d \ln(2m)$ as required.

Application: Unrealizable (Agnostic) Learning

- We are given a training set $\{(x_1, c(x_1)), \ldots, (x_m, c(x_m))\}$, and a concept class $\mathcal{C}$
- No hypothesis in the concept class $\mathcal{C}$ is consistent with all the training set ($c \notin \mathcal{C}$).
- Relaxed goal: Let c be the correct concept. Find $c' \in \mathcal{C}$ such that

$$\Pr_{\mathcal{D}}(c'(x) \neq c(x)) \leq \inf_{h \in \mathcal{C}} \Pr_{\mathcal{D}}(h(x) \neq c(x)) + \epsilon.$$

- An $\epsilon/2$-sample of the range space $(X, \Delta(c, c'))$ gives enough information to identify an hypothesis that is within ϵ of the best hypothesis in the concept class.
- The range spaces $(X, \mathcal{C})$ and $(X, \Delta(c, c'))$ have the same VC-dimension.

Uniform Convergence

Definition

A range space $(X, \mathcal{R})$ has the *uniform convergence property* *if for every* $\epsilon, \delta > 0$ *there is a sample size* $m = m(\epsilon, \delta)$ *such that for every distribution* $\mathcal{D}$ *over* X, *if* S *is a random sample from* $\mathcal{D}$ *of size* m *then, with probability at least* $1 - \delta$, S *is an* ϵ*-sample for* X *with respect to* $\mathcal{D}$.

Theorem

The following three conditions are equivalent:

1. *A concept class* $\mathcal{C}$ *over a domain* X *is agnostic PAC learnable.*
2. *The range space* $(X, \mathcal{C})$ *has the uniform convergence property.*
3. *The range space* $(X, \mathcal{C})$ *has a finite VC dimension.*

Is VC-Dimension "Just a Theory"?

Two issues:

- Hard to prove an efficient bound on VC-dimension
- VC-dimension is a "worst case" bound

A quick example:

- Very easy to compute bound on VC-dimension
- Better than union bound
- Not a machine learning problem

Frequent Itemsets Mining (FIM)

Frequent Itemsets Mining: classic data mining problem with many applications. Settings:

Dataset $\mathcal{D}$
bread, milk
bread
milk, eggs
bread, milk, apple
bread, milk, eggs

Each line is a transaction, made of items from an alphabet $\mathcal{I}$

An itemset is a subset of $\mathcal{I}$. E.g., the itemset $\{\text{bread,milk}\}$

The frequency $f_{\mathcal{D}}(A)$ of $A \subseteq \mathcal{I}$ in $\mathcal{D}$ is the fraction of transactions of $\mathcal{D}$ that A is a subset of.

E.g., $f_{\mathcal{D}}(\{\text{bread,milk}\}) = 3/5 = 0.6$

Problem: Frequent Itemsets Mining (FIM)

Given $\theta \in [0,1]$ find (i.e., mine) all itemsets $A \subseteq \mathcal{I}$ with $f_{\mathcal{D}}(A) \geq \theta$

I.e., compute the set $\mathrm{FI}(\mathcal{D},\theta) = \{A \subseteq \mathcal{I} \ : \ f_{\mathcal{D}}(A) \geq \theta\}$

FI mining algorithms (Apriori, FP-Growth, ...) require significant computation time and space ($\geq$ quadratic in number of transactions). What can be done with a sample?

What can we get with a Union Bound?

For any itemset A, the number of transactions that include A is distributed

$$|\mathcal{S}|f_{\mathcal{S}}(A) \sim \mathit{Binomial}(|\mathcal{S}|, f_{\mathcal{D}}(A))$$

Applying Chernoff bound

$$\Pr(|f_{\mathcal{S}}(A) - f_{\mathcal{D}}(A)| > \varepsilon/2) \leq 2e^{-|\mathcal{S}|\varepsilon^2/12}$$

We then apply the union bound over all the itemsets to obtain uniform convergence

There are $2^{|\mathcal{I}|}$ itemsets, a priori. We need

$$2e^{-|\mathcal{S}|\varepsilon^2/12} \leq \delta/2^{|\mathcal{I}|}$$

Thus

$$|\mathcal{S}| \geq \frac{12}{\varepsilon^2}\left(|\mathcal{I}| + \ln 2 + \ln\frac{1}{\delta}\right)$$

Assume that we have a bound ℓ on the maximum transaction size.

There are $\sum_{i \leq \ell} \binom{|\mathcal{I}|}{i} \leq |\mathcal{I}|^{\ell}$ possible itemsets. We need

$$2e^{-|\mathcal{S}|\varepsilon^2/12} \leq \delta/|\mathcal{I}|^{\ell}$$

Thus,

$$|\mathcal{S}| \geq \frac{12}{\varepsilon^2}\left(\ell \log |\mathcal{I}| + \ln 2 + \ln \frac{1}{\delta}\right)$$

The sample size still depends on $|\mathcal{I}|$, which can be very large - all products sold by Amazon, all the pages on the Web, ...

Can we have a smaller sample size?

How do we get a smaller sample size?

[Riondato and U. 2014, 2015]: Let's use VC-dimension!

- The domain is the dataset $\mathcal{D}$ (set of transactions)
- For each itemset $A \subseteq 2^{\mathcal{I}}$ we have the set of transactions that contain A

$$\mathcal{T}_A = \{\tau \in \mathcal{D} \ : \ A \subseteq \tau\}$$

- We we need to estimate the probabilities (sizes) of all ranges in the range space

$$(\mathcal{D}, \{\mathcal{T}_A, A \subseteq 2^{\mathcal{I}}\})$$

We need an efficient-to-compute upper bound to the VC-dimension

How do we bound the VC-dimension?

Definition

The d-index of a dataset $\mathcal{D}$ is the maximum integer d such that $\mathcal{D}$ contains at least d different transactions with at least d items

Example: The following dataset has d-index 3

bread	beer	milk	coffee
chips	coke	pasta	
bread	coke	chips	
milk	coffee		
pasta	milk		

It can be computed easily with a single scan of the dataset

Theorem

The VC-dimension of $\mathcal{D}$ is bounded by the d-index of $\mathcal{D}$

How do we prove the bound?

Theorem: The VC-dimension is less or equal to the d-index d of $\mathcal{D}$

Proof:

- Let $\ell > d$ and assume it is possible shatter a set $T \subseteq \mathcal{D}$ with $|T| = \ell$.
- Then any $\tau \in T$ appears in at least $2^{\ell-1}$ ranges $\mathcal{T}_A$ (there are $2^{\ell-1}$ subsets of T containing τ)
- But any τ only appears in the ranges $\mathcal{T}_A$ such that $A \subseteq \tau$. So it appears in $2^{|\tau|} - 1$ ranges
- From the definition of d, T must contain a transaction τ^* of length $|\tau^*| < \ell$
- This implies $2^{|\tau^*|} - 1 < 2^{\ell-1}$, so τ^* can not appear in $2^{\ell-1}$ ranges
- Then T can not be shattered.

How good is the bound?

Definition

The d-index of a dataset $\mathcal{D}$ is the maximum integer d such that $\mathcal{D}$ contains at least d different transactions with at least d items

If all transactions have exactly ℓ elements , then $d = \ell$.

If we have n transactions, the largest transaction has ℓ elements, and the number of elements in a transaction follows a power law distribution

$$Pr(X \geq x) \sim cx^{-\alpha}, \quad \text{for } \alpha > 1,$$

then d satisfies $Pr(X \geq d) \sim \frac{d}{n}$, and ℓ satisfies $Pr(X \geq \ell) \sim \frac{1}{n}$, which gives,

$$d \sim \ell^{\frac{\alpha}{1+\alpha}}$$

Frequent Itemset Estimation Using VC-dimension

The VC-dimension is bounded by the maximum d such that $\mathcal{D}$ contains at least d different transactions with at least d items.

Sample size

$$|S| = O\left(\frac{d}{\epsilon^2}\ln\frac{d}{\epsilon} + \frac{1}{\epsilon^2}\ln\frac{1}{\delta}\right)$$

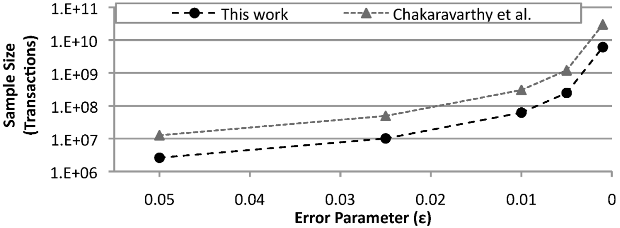


Figure: Frequent itemsets: Sample size based on VC-dimension vs. union bound